DOCUMENT RESUME

ED 129 853 TM 005 473

AUTHOR	Scheuneman, Janice
TITLE	Validating a Procedure for Assessing Bias in Test Items in the Absence of an Outside Criterion.
PUB DATE	[Apr 76]
NOTE	12p.; Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, April 19-23, 1976)
EDRS PRICE	MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS	*Evaluation Criteria; *Item Analysis; Response Style (Tests); Statistical Analysis; *Test Bias; Test Construction; Test Validity; Validity
IDENTIFIERS	Metropolitan Readiness Tests

ABSTRACT

Test bias has traditionally been defined in terms of an outside criterion measure of the performance being predicted by the test. In test construction, where criterion-related validity data are usually not collected until after the test is completed, assessment of bias in the absence of outside criteria had become a vital issue. Here, an unbiased test item is defined as one in which, for persons with the same ability in the areas being measured, the probability of a correct response on the item is the same regardless of the population group membership of the individual. The total score on a test or subtest containing the item can be used to define groups of persons having the same ability. Once the ability groups have been defined, a modified chi square procedure is used to evaluate each item in the test for possible bias. While hypotheses suggested by such an evaluation should be investigated further before making conclusive statements concerning the source of bias, results reported in this study support the validity of the method as a procedure for assessing bias when outside criterion measures are unavailable. (BW)

VALIDATING A PROCEDURE FOR ASSESSING BIAS IN TEST ITEMS[1]
IN THE ABSENCE OF AN OUTSIDE CRITERION

Janice Scheuneman

The Psychological Corporation

During the past few years the problem of bias in testing has become an increasingly important issue. In most of the research which has been done, bias refers to the fair use of tests and has thus been defined in terms of an outside criterion measure of the performance being predicted by the test. Recently, however, there has been growing interest is assessing bias when such criteria are not available. In test construction in particular, where criterion-related validity data are usually not collected until after the test is completed, assessment of bias in the absence of outside criteria has become a vital issue. If tests are to be built which may someday prove to be unbiased in use, it is important to identify potentially biased items during the construction process when test content is still flexible and items may still be modified or replaced. In addition, the identification of such items is a first step in isolating sources of bias in the test content, information which is potentially useful to researchers in other areas interested in population group differences as well as for test construction purposes in the future.

Procedure

In the method discussed in this paper, an unbiased item is defined as one in which, for persons with the same ability in the areas being measured, the probability of a correct response on the item is the same regardless of the population group membership of the individual. In cases where no outside criterion measures of ability are available, the total score on a test or subtest containing the item can be used to define groups of persons having the same ability. Assuming that the test is reasonably valid and reliable and is homogeneous with respect to the ability being measured, the definition can be restated as follows: An item is unbiased if, for all individuals belonging to the same ability group as defined by the total score on the test or subtest containing the item, the proportion of individuals getting the item correct is the same for each population group being considered. Once the ability groups have been defined, a modified chi square procedure is used to evaluate each item in the test for possible bias.

Table 1 gives a computational example of the procedure. This procedure differs from the conventional chi square test primarily in the computation of the expected frequencies. The first column gives the score ranges which were selected for this item. In general, score ranges have been selected by dividing the distribution of correct responses approximately at the quartiles or quintiles for the smaller-sized group. The next set of columns are the frequency distributions of scores both within and across the two population groups. The next columns give the obtained frequencies--the number of children within each population/

score range group who got the item correct. The next column, proportion correct, is computed across population groups. Within each score range group the total number of correct responses is divided by the total number of children scoring in that range. This proportion is then used to obtain the expected frequency for each cell. According to the definition, if the item is unbiased, the proportion of correct responses should be the same regardless of population group membership. Hence, the expected number of correct responses is obtained by multiplying the proportion p by the number of children in the population group who scored in that range. In this item, a comparison of the obtained and expected number of correct responses will quickly show that Black children are doing consistently more poorly than expected on this item. The item would probably be considered biased depending on the cut-off point chosen.

Developed initially as part of the item analysis program for the Metropolitan Readiness Tests, the procedure was used to screen the item pool for items which were potentially biased (Scheuneman, 1975). As a rapid screening device it proved quite satisfactory. The method is computationally simple and permits easy establishment of a decision rule for rejection of items. It is not necessary to assume that the groups are representative of their respective populations, nor are any normality assumptions required. Very easy items can be evaluated without difficulty, although very difficult items present problems. Any item where one population group produces fewer than ten correct responses cannot be evaluated at all. Fairly large samples are required for the method, probably about 100 per population group.

If the method presented here is a valid procedure, examination of the items selected as biased should yield further information about possible sources of the bias. During the 1975 standardization of the Metropolitan Readiness Tests, a study designed to produce norms for large cities was conducted, with 11 cities from across the country participating (Psych. Corp., 1976). Items in Level II Form P of the test were analyzed for bias using data collected during this program. Level II consists of a total of 97 items from four "skill areas"--Auditory, Visual, Language, and Quantitative, each of which is made up of two subtests. The sample consisted of 4441 First Grade children of whom 1653 were identified as White, 1502 as Black, 470 Mexican American , 161 Puerto Rican and 123 Oriental. A total of 532 children belonged to other population groups or were unidentified and were not included in the analysis. The items were screened by using a 5 x r chi square, where there were five population groups and r score range groups, r ranging from two to five. Items found to be biased were examined further using tests with two, three, or four population groups at a time as seemed indicated in order to get at the patterns of differences between the groups.

Results

From the 97 items, 34 items were found to be biased using the 2 five population groups together. With five of these items, significance appeared to result from the particular choice of interval, that is, when the score intervals were changed, the results were no longer significant. Nine items, although consistently showing bias with different intervals and with different combinations of population groups, revealed no clear pattern of results. Another five items showed few significant differences when the population groups were tested two at a time, but instead appeared to rank the groups by performance, with significance resulting only between the extreme groups. With the remaining 14 items clear patterns of bias were found, either for or against one or two population groups.

The items in the Auditory area yielded some of the most easily interpretable results and nicest examples. In the Beginning Consonants Test, for example, Oriental children were found to have undue difficulty discriminating between an L and an R. In the other subtest, Sound-Letter Correspondence, children were asked to select the letter which corresponds to the beginning sound of a word which is pictured on the test and named by the teacher. One of these items, in which the stimulus word was dog, was found to be biased against both of the Spanish-speaking groups. On investigation, it was noted that the Spanish word for dog was perro and that the distractors included a p.

In the Visual Skill area, eight items were found to be biased. Although two or three of these seemed clearly biased in favor of Oriental children, generally the patterns of differences were not clear. When examined for content, however, five of the biased items were found to involve artificial letters or letter-like shapes although only eight items involving these letters were included in the 26 item test. This finding is at least suggestive of a possible source of bias in the test which would warrant further investigation.

The results for the Visual tests were further complicated, however, in that the lower range was so much restricted for the Oriental children. No Oriental child scored low on this test with the result that the score range for the lowest group was unusually wide, possibly covering as many as 14 or 15 points. (A lower limit of at least ten correct responses per cell was observed in all cases.) With only a few children at the top of the interval for one group versus a large number of children acrosu a wide range of scores in the other groups, the assumption of equal ability within the score range is no longer very tenable. (A similar distribution at the upper end of the scale, however, does not appear to create problems. Within the top scoring groups, the differences between expected and obtained frequencies is seldom very large, even though the upper range of scores may vary widely among the population groups.)

The Language area consists of only 18 items, of which 11 were found to be biased, seven of them beyond the .05 level.[3] Not too surprisingly, most of these items were found to be biased against Spanish-speaking or Oriental children, with the biased items involving more complex grammatical structures than the unbiased items.

In the item analysis most of the biased items were in the Quantitative area, but in constructing the final version of the Metropolitan Readiness Tests at Level II, the quantitative items were broken into two subtests, Quantitative Concepts and Quantitative Operations. The Quantitative Concepts Test contains items measuring concepts such as part-whole relations and one-to one-correspondence, some spatial perception items; and some simple figure analogies. The Quantitative Operations Test primarily contains fairly straightforward counting and simple computation problems. Looking at the two subtests separately, five items from the nine item Quantitative Concepts Test (55% of the items) and four from the 15 item Quantitative Operations Test (27% of the items) were found to be biased.

When the results from the Quantitative Concepts and the School Language tests are examined together, a pattern appears which suggests that Black children have trouble with terms such as "fewer," "closer," "larger." This pattern was discernible in the item analysis data, but not so clearly visible as in this sample where all four items concerning such terms were found to be biased against Black children.

In the item analysis program potentially biased items were usually discarded, but for a number of reasons there were too few remaining items in some areas and five items which were apparently biased, but otherwise satisfactory, were included in this form of the test. Of these, three again appeared as biased, while the stem of a fourth item had been extensively revised in an effort to make the task involved clearer--possibly removing the source of the bias in the original item.

Summary and Conclusion

Any method for assessing bias which uses only information contained within the test is open to criticism concerning the validity of the procedure. Using internal statistics alone, it is not possible to determine if the method is in fact isolating items which are biased or simply selecting items more or less at random. While some false positives are to be expected, examination of the content of the items should reveal some specific item content or pattern of content which is interpretable in light of knowledge beyond that yielded by the test. While hypotheses suggested by such an examination should be investigated further before making conclusive statements concerning the source of bias, results such as those reported in this study support the validity of the method as a procedure for assessing bias when outside criterion measures are unavailable.

References

The Psychological Corporation. Norms tables for large city school systems (MRT Research Research No. 3). New York: Author, 1976.

Scheuneman, J. A new method of assessing bias in test items. Paper presented at the meeting of the American Educational Research Association, Washington, D. C., April 1975.

Footnotes

1. This paper is a slightly modified version of a paper presented at the meeting of the American Educational Research Association as part of a symposium entitled "The Assessment of Bias in the Absence of an Outside Criterion," San Francisco, April 1976.

2. In determining if an item was biased or unbiased, a standard chi square table was entered with the obtained chi square value and (r-1)(k-1) degrees of freedom where r is the number of score groups and k is the number of population groups. If the probability of the obtained chi square was read to be less than .30, the item was termed biased. It should be noted that .30 is not the probability of rejecting an unbiased item in the hypothesis testing sense. It is an arbitrarily selected cutting point which serves to isolate those items which are most likely to be biased by the definition given here. This point was selected during the item analysis program for eliminating potentially biased items with the idea that it was better to reject an unbiased item than to retain a biased one, while still not eliminating so many items that the item pool would become too small. The .30 cutting point seemed to strike a good balance and was retained for this study. Further work is still needed to determine the various statistical properties of the test.

3. Again the .05 level refers to the cutoff points when using the chi square tables rather than the probability of rejecting an unbiased item.

Table 1

Example of the Computation of Chi Square

for one item

Total Score on Subtest	Number with Scores in each Range			Obtained frequencies (Number with item correct)			Proportion Correct	Expected frequencies	
	Black	White	Total	Black	White	Total	(T_o/T)	Black (p B)	White (p W)
	B	W	T	B_o	W_o	T_o	p	B_e	W_e
12-15	12	350	362	10	320	330	.9116	10.94	319.06
10-11	34	152	186	17	104	121	.6505	22.12	98.88
8-9	24	66	90	8	40	48	.5333	12.80	35.20
3-7	33	47	80	6	15	21	.2625	8.66	12.34

$$\chi^2 = \sum \frac{(B_e - B_o)^2}{B_e} + \sum \frac{(W_e - W_o)^2}{W_e} = 5.317$$